DOCUMENT RESUME

ED 283 842 TM 870 355

AUTHOR Mills, Craig N.; Melican, Gerald J.
TITLE A Preliminary Investigation of Three Compromise Methods for Establishing Cut-Off Scores.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-87-14
PUB DATE Mar 87
NOTE 28p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Standards; Certification; Comparative Analysis; *Cutting Scores; *Difficulty Level; *Judges; Knowledge Level; Licensing Examinations (Professions); *Mastery Tests; Statistical Distributions; Statistical Studies; *Test Items
IDENTIFIERS Angoff Methods; *Compromise Model (Hofstee)

ABSTRACT

The study compares three methods for establishing cut-off scores that effect a compromise between absolute cut-offs based on item difficulty and relative cut-offs based on expected passing rates. Each method coordinates these two types of information differently. The Beuk method obtains judges' estimates of an absolute cut-off and an expected passing rate, and constructs a cutting line whose slope is the ratio of the absolute and relative standard deviations and which passes through the point of mean absolute/relative cut-off. The judges can be either test-oriented or examinee-oriented depending on whether they show greater agreement (small standard deviations) on the absolute or relative cut-offs. The Hofstee method draws a cutting line through two extreme points: (1) maximum cut-off, minimum failure point; and (2) minimum cut-off, maximum failure point. The DeGruijter method is similar to the Beuk method, but uses confidence estimates for the absolute and relative cut-offs to define a criterion ellipse. These methods were applied to two tests from a certification program. Judges rated item difficulty by the Angoff method and estimated a desirable passing rate. All three compromise methods brought the cut-off two points below the absolute level, in line with an acceptable passing rate. This study suggests that further research into all three of the compromise methods is needed. (LPG)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED283842

RR-87-14

BEST COPY AVAILABLE

RESEARCH REPORT

A PRELIMINARY INVESTIGATION OF THREE COMPROMISE METHODS FOR ESTABLISHING CUT-OFF SCORES

Craig N. Mills
Gerald J. Melican

TM 870 355

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Educational Testing Service
Princeton, New Jersey
March 1987

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
H. C. Weitzman Miller
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A Preliminary Investigation of Three Compromise Methods for Establishing Cut-Off Scores

Craig N. Mills

Gerald J. Melican

Educational Testing Service

Copyright © 1987. Educational Testing Service. All rights reserved.

January 26, 1987

A Preliminary Investigation of Three Compromise Methods for Establishing Cut-off Scores

Craig N. Mills
and
Gerald J. Melican

The determination of passing points for tests used to make dichotomous decisions about individuals has been an area of concern for some time. Many methods have been proposed and used to establish cut-off scores. These methods have been grouped into two major classes: those that include judgments about examinees and those that require judgments about items (Livingston & Zieky, 1982). Most cut-off score studies in the area of licensure and certification have fallen into the latter category; that is, methods that require judgments about items have been predominant. The most widely used methods have been the Nedelsky, Angoff, and Ebel methods.

One problem associated with the use of item judgment methods has been the low to moderate relationship of raters' perceptions of item difficulty to the actual item difficulty. Although different training procedures have been used with varying degrees of success (e.g. Bejar, 1983, Thorndike, 1982, Lorge and Kruglov, 1953), the accuracy of judges' estimates of item difficulty is generally regarded as less than adequate. As a result, cut-off score studies are typically viewed as tentative and the cut-off scores obtained from standard setting studies are routinely adjusted. Traditionally, cut-off scores have been adjusted by lowering (or raising) them by a multiple of the standard error of measurement or the conditional standard error of measurement

(Lord, 1984) at the cut-off point. The arguments used to make adjustments using the standard error of measurement are usually made in terms of the type of error least objectionable to the sponsor of the test. For example, if the sponsor feels that it is not as harmful to pass a candidate whose ability is actually lower than the cut-off as to fail an able examinee who scored poorly, the cut-off may be lowered.

A decision to adjust a cut-off based on an evaluation of the relative costs of false positive and false negative errors may be appropriate, however, the passing rates associated with the alternative cut-off scores are often inappropriately considered as part of this evaluation. The passing rates are not directly related to the numbers of errors of either type (false positive or false negative) possible for the various alternative cuts and, as such, the passing rates are often reviewed in an inappropriate manner, a manner more consistent with the setting of the cut-off score to pass a fixed percentage of examinees. This is not to say, however, that the client's or judges' estimates of the percent of examinees in the entry population that should pass is not important information. To the contrary, these estimates are important collateral information because they are generally based on solid observation of the examinee population. The manner in which this information is to be incorporated into the decision regarding the placement of the cut-off is an issue to be considered.

Three methods of adjusting absolute cut-off scores that incorporate

expected passing percentages have recently been proposed in two contributions to the literature (Beuk, 1984, DeGruijter, 1985). These methods purport to provide compromises between an absolute cut-off score (a cut-off based solely on the examinee's performance on the test) and a relative one (a cut-off based solely upon the examinee's ranking in some group). These compromise methods, if they yield acceptable cut-offs, may offer standardized ways of adjusting cut-off scores. This paper is written to explain the methods and to present preliminary data differences in the application of them.

Compromise Methods

In this paper, three compromise methods will be explained and illustrated. The word "compromise" is used to mean methods that incorporate both judges' estimates of an absolute cut-off and their estimates of the pass rates in establishing a cut-off score.

The Beuk Method

Beuk (1984) presented a method in which the cut-off is adjusted as a function of the degree to which the judge group is identified as "test oriented" or "examinee oriented." In order to use the Beuk method, two pieces of information are required from each judge: a cut-off score and an estimate of the percent of the candidate group that should pass. The score point

associated with the estimate of percent pass is referred to as the relative cut-off score, i.e. the cut-off score that would result in the pass rate predicted by the judges. Additionally, a distribution of scores is required.

In order to determine whether a group of judges is test or examinee oriented, one compares the standard deviations of their item ratings and their percent pass ratings. If a smaller standard deviation is noted for the absolute cut-off score data than for the expected percent passing data, the group of judges is considered to be test oriented. If the reverse is true, the group is called examinee oriented. The ratio of the standard deviations of the ratings on the absolute and relative dimensions is then used to adjust the cut-off as described below:

-a cumulative distribution of the percent of the examinee group falling below each possible cut-off score (each test score is a possible cut-off) is plotted.

-a point corresponding to the mean cut-off score and the mean expected pass rate is plotted on the graph.

-a line with a slope equal to the ratio of the standard deviations (i.e. the standard deviation of the judges' expected pass rates divided by the standard deviation of their cut-off scores) is drawn through the point representing the two means.

The point where the line passes through the cumulative distribution of scores is the adjusted cut-off. In this report, the Beuk method is demonstrated in terms of a cumulative distribution of percent fail to maintain continuity with DeGruijter (1985) who describes all three methods in terms of percent fail.

The adjusted cut-off score via the Beuk method will be adjusted in relation to the two suggested cut-off scores (the absolute and relative cuts). If the standard deviation for the absolute cut-off is the smaller of the two standard deviations, the adjusted cut-off will be closer to the absolute cut-off score than to the expected percent passing cut-off score. In this way, the Beuk method favors the cut-off score technique for which the judges show the stronger agreement (smaller standard deviation) and the adjusted cut-off score will tend toward one or the other of the cut-off scores as the ratio of the standard deviations departs from unity.

The Hofstee Method

Hofstee's method may be used as a method for setting a cut-off score in addition to its use as a method for adjusting one. Four judgments are required from each judge using Hofstee's method:

-an estimate of the maximum acceptable value of the cut-off (C_{max}); the cut-off should not be higher than this point even if every examinee passes the test,

-an estimate of the minimum acceptable value of the cut-off (C_{min}); the cut-off should not be lower than this point even if no examinee passes,

-an estimate of the minimum acceptable failure rate (F_{min}), and

-an estimate of the maximum acceptable failure rate (F_{max}).

Two points are plotted: the maximum cut-off, minimum failure point and the minimum cut-off, maximum failure point. The line-segment that connects the points establishes acceptable combinations of cut-off scores and failure rates. A cumulative distribution of percent failing is plotted on the same graph. The point where the cumulative distribution intersects the line-segment becomes the cut-off.

The Hofstee method evaluates "worst case" possibilities: "Based on the information provided in the responses to the cut-off score and failure rate questions, we would be willing to accept a cut-off score as high as C_{max} provided the failure rate did not exceed F_{max}. Further, we would accept a cut-off score as low as C_{min} provided the failure rate was at least F_{min}."

The points on the line-segment connecting the extremes C_{max}, F_{min} and C_{min}, F_{max} represent the acceptable alternative combinations of cut-off scores and passing rates. The point on this line-segment that coincides with the ogive is the point where the judges are in agreement with the observed data.

The possibility exists that the line-segment established via the Hofstee method may not cross the ogive indicating that the judges' estimates of the range of possible cut-off scores or the range of possible passing rates (or both) were inconsistent with the performance of the examinees. This is most likely when the judges are in strong agreement about one or both of the extreme cut-off scores and the range between the extremes is small. Decisions need to be made about how to proceed if this disagreement between the judges' estimates and observed data should occur. It should also be noted that the line-segment drawn using the Hofstee method need not intersect the point representing the mean cut-off score and mean percent pass obtained using the Beuk method. The Beuk method uses means and standard deviations, while the Hofstee method uses only extreme values.

The DeGruijter Method

The DeGruijter method is similar to the Beuk method. Each judge provides a cut-off score and an expected pass rate. Additionally, however, judges must provide estimates of their confidence in their ratings on both the absolute and relative dimensions (or the ratio of their confidence in those ratings).

The DeGruijter method identifies the one member of a family of ellipses that just touches the ogive. The family of ellipses is defined by the equation:

$$d^2 = r^2(c_o - c_i)^2 + (f_o - f_i)^2 \qquad (1)$$

where d is half of the length of the ellipse in the vertical direction,

r is the ratio of the judges' uncertainty with respect to the true value of f_i to their uncertainty about the true value of c_i (u_f/u_c),

c_o is an observed cut-off (a test score),

c_i is the ideal cut-off (from the cut-off score study),

f_o is the observed failure rate at c_o, and

f_i is the ideal failure rate (from the cut-off score study).

The values of c_o and f_o that yield the smallest value of d define the one ellipse in the family of ellipses from Formula 1 that just touches the ogive. C_o is then taken as the adjusted cut-off that provides the best compromise

between the absolute and relative cut-off scores. The point c_o, f_o does not have to be a whole number, but the equation may be solved using whole numbers.

The ratio of the uncertainty estimates determines the amount of compromise required along the cut-off score and failing percent continua. A ratio greater than 1.0 (the judges were more uncertain about the estimates of the failing rates than they were about the estimates of the cut-off scores) will result in an ellipse in which the vertical axis is the major (longer) axis. This type of ellipse will tend to result in a larger discrepancy between the adjusted failure rate and the judges' estimated failure rate than between the adjusted cut-off score and the judges' estimated cut-off score. A ratio less than 1.0 will have the opposite effect; the horizontal axis will be the major axis.

Procedures and Instruments

Two tests were included in the study. Both tests are included in the same certification testing program, but cover different content areas. Separate panels of judges were convened to rate the two tests. The Angoff method was used to obtain ratings of item difficulty for the items in each test. Prior to the rating of test items, each judge responded to the following question: If the test were a perfect instrument for measuring exactly what candidates knew, what percentage of candidates taking the test

should pass? Each test contained 34 four-option multiple-choice items. Descriptive information about the tests is shown in Table 1.

The data collected were directly applicable to the Beuk method, but the the collection of the additional data required for the DeGruijter and Hofstee method was not included in the study design. To demonstrate the DeGruijter method, the ratio of the standard deviations of the absolute and relative cut-offs was used as a proxy for the ratio of the uncertainty estimates. The outcome of this decision was that the DeGruijter results are very similar to the Beuk results. To demonstrate the Hofstee method, the highest cut-off resulting from the Angoff data collection was used as the maximum acceptable cut-off, the lowest Angoff cut-off was taken as the minimum acceptable cut-off and the maximum and minimum failure rates were set from the values obtained from the judges' responses to the expected percent pass question.

Results

The Angoff cut-off scores and the expected pass rates are shown in Table 2. The group rating Test 1 can be described as "test oriented". The standard deviation of the absolute cut-off scores is approximately one-third as large as the standard deviation of the expected pass rates (6.13 and 16.71 respectively). The standard deviations for the group rating the second test

were much more similar (2.34 for the absolute cut-off and 1.91 for the passing rates).

Tables 3 and 4 show the Beuk, DeGruijter, and Hofstee results for Test 1 and Test 2 respectively. The tables contain data only for the range of scores within which the cut-off could conceivably be expected to lie. The columns in the tables were derived as follows:

Cut-off score - each possible test score was used as a potential cut-off in the calculations

% Below - the percentage of the examinee group below the raw score

Beuk Y - the value Y in the equation Y = aX + b where a is the slope (ratio of the standard deviations multiplied by -1), b is the intercept and X is the raw score expressed as a percent

Beuk Diff - the Beuk Y - % Below: the discrepancy between the observed failure rate and the Beuk value

DeGruijter d - explained previously under the DeGruijter method

Hofstee Y - the value Y in the equation Y = aX + b where a is the slope (determined from the two given points), b is the intercept and X is the raw score expressed as a percent. The line has end points as described in the Hofstee section of the paper.

Hofstee Diff - the Hofstee Y - % Below: the discrepancy between the observed failure rate and the Hofstee value

The suggested cut-off can be determined by locating the number with the smallest absolute value in the difference column for Beuk and Hofstee and in the d column for DeGruijter. For both tests, the application of each method resulted in a two point drop in the cut-off score. The initial cut-off, based on the judges' Angoff estimates was 22 items correct. The compromise methods reduced the cut-off to 20.

These results are depicted graphically in Figures 1 and 2 for Test 1 and Test 2, respectively. In Figure 1, the independence of the Hofstee line from the results of the Angoff method shows clearly. Although the same number correct score is obtained when the results are rounded, the Hofstee line does not include the point defined by the two means. The DeGruijter ellipse shown in Figure 1 is elongated along the vertical axis. As noted previously, this

represents the greater spread of estimates for the relative cut-off than for the absolute cut-off.

Figure 2 shows a DeGruijter ellipse that is slightly elongated along the horizontal dimension. The uncertainty ratio in this case was less than unity since there was greater variance in the absolute than the relative cut-off. Also, the difference in the standard deviations was less for Test 2 than for Test 1, resulting in a more circular ellipse. This figure shows a very short Hofstee line-segment. Judges were in close agreement as to the relative cut-off. Their agreement was not, however, sensitive to the data. The result was a short line-segment that did not cross the cumulative distribution. Therefore, for these data the Hofstee method did not yield a cut-off score.

In order to better illustrate the DeGruijter method, three ellipses are shown for Test 1 and Test 2 in Figures 3 and 4, respectively. The three ellipses shown in each figure are: the ellipse yielding the compromise cut-off and the ellipses based on the data one score point above or one point below that value. In Figure 3, the ellipse drawn using the data from the compromise cut-off value actually touches the cumulative frequency distribution at that point, while the other two ellipses are clearly larger and intersect the distribution at two points. In Figure 4, it can be seen that the ideal compromise would be a non-integer value since the compromise ellipse actually intersects the cumulative distribution at two points.

However, the smallest ellipse drawn using integer values is used as the compromise cut-off since the score scale is an integer scale.

These two figures demonstrate how the shape of the ellipses and the resulting modification of the cut-off are affected by the magnitude of the uncertainty ratio. In Figure 3, the ellipses are elongated in the vertical axis. The uncertainty ratio for that test was 2.73. For this test, the judges estimated the fail rate to be 33 percent and the cut-off score to be 65 percent (i.e. 22 items out of 34). The adjusted fail rate was 45 percent and the adjusted cut-off score was 59 (i.e. 20 items out of 34). The difference between the original and adjusted fail rates was 12 percent, compared to a difference of 6 percent for the absolute cut-off scores, consistent with an ellipse that has the vertical axis as its major axis. Figure 4, on the other hand, was generated with an uncertainty ratio of 0.82 from Test 2 and indicates an elongation along the horizontal axis. As expected with this type of ellipse, the difference between the ideal and compromise cut-offs (65 and 59 percent respectively) is greater than the difference between the ideal and adjusted fail rates (29 percent in both cases).

Discussion

An evaluation of the usefulness of the three compromise methods can be made by considering the effect they have on the passing rate. For Test 1, implementation of the judges' initial cut-off (a raw score of 22 or 65 percent

correct) would have provided a 33 percent passing rate. The compromise methods raised that rate to 55 percent by lowering the cut-off to a raw score of 20 (59 percent correct). The 55 percent pass rate following the adjustment is within twelve percent of the average desired passing percent (67 percent) for that test. One judge rated the expected passing rate for this test as 33 percent. Although, in practice one would not remove a judge simply on the basis of a large discrepancy, it is instructive to note the effect of that judge. Without the data for the sixth judge, the expected raw score cut-off remains 22. The expected pass rate is raised from 67 to 72.5 percent. The standard deviations (in terms of percentages) for the cut-off score and the pass rate are 6.72 and 8.22 respectively. These data would result in a three point drop in the cut-off and a final pass rate of 65 percent.

For Test 2, the desired passing percentage was 71. The initial cut-off (a raw score of 22) resulted in a 50 percent pass rate. Following application of the compromise methods, the passing rate at the resulting raw score cut-off of 20 was 71 percent.

The results of this preliminary investigation are encouraging. Although the methods provided the same results, this is probably due to the dependency of the Hofstee and DeGruijter results on data collected for the Beuk method. Nonetheless, the methods all provided compromise cuts resulting in passing rates that were reasonably close to the rates specified by the judges.

By inspecting Tables 3 and 4, it can be seen that the portion of the score scale containing the original and compromise cuts is an area in which many examinees lie. Thus, the effect of a change in the cut-off of even a single point on the overall pass rate is substantial. Had the cut-off been in a different portion of the distribution, the results may not have been as striking.

This study suggests that further research into all three of the compromise methods is needed. A study which compares the methods when data have been collected specifically for each method may better clarify differences among the methods than this study which was intended primarily to demonstrate differences in the ways the data are treated. Additional studies will be also required to investigate the sensitivity of the methods to various combinations of score distributions and placement of the cut-off within those distributions. As further research is conducted, important information about the conceptual and practical attractiveness of the methods to client groups will also become available. Consideration should be given to research concerning more effective methods of acquiring passing rate data, including what information should be provided to judges and how to establish uncertainty estimates.

Table 1
Descriptive Information for the Two Tests

	Test 1	Test 2
Number of Items	34	34
Number of Examinees	784	228
Mean	19.81	21.32
Standard Deviation	4.20	4.11
Skewness	-.19	.89
KR-20	.62	.60
SEM	2.59	2.60
SEM (conditional)*	2.67	2.64

*Lord, F.M. (1984)

Table 2
Estimates of Absolute Cut-off Scores and Expected Pass Rates

	Test 1		Test 2	
Judge	Cut-off Score (raw)	Expected Fail Rate	Cut-off Score (raw)	Expected Fail Rate
1	18.20	30	20.85	30
2	24.40	35	20.85	30
3	20.95	30	22.95	28
4	20.70	20	22.00	30
5	22.05	35	20.95	30
6	21.80	67	21.85	25
7	23.90	15	21.10	30
Mean	21.71 (63.87%)	33	21.51 (63.26%)	29
SD	2.08 (6.13%)	16.71	0.79 (2.34%)	1.91

Table 3
Results of the Application of the Beuk, DeGruijter, and Hofstee Methods for Adjusting Cut-off Scores for Test 1

Raw Score	Actual Percent Below	Beuk Values Y	Beuk Values Diff	DeGruijter d	Hofstee Values Y	Hofstee Values Diff
15	10	87	77	58		
16	14	79	65	49		
17	21	71	50	40		
18	28	63	35	30	69	41
19	35	55	20	22	60	25
20	45	47	2	18	52	7
21	57	39	-18	25	44	-14
22	67	31	-36	34	35	-32
23	74	23	-52	42	27	-48
24	81	15	-66	51	18	-63
25	88	7	-81	60		
26	92	-1	-93	68		
27	95	-9	-104	75		
28	97	-17	-114	81		
29	98	-25	-124	88		
30	99	-33	-132	93		

Table 4
Results of the Application of the Beuk, DeGruijter, and Hofstee Methods for Adjusting Cut-off Scores for Test 2

Raw Score	Actual Percent Below	Beuk Values Y	Beuk Values Diff	DeGruijter d	Hofstee Values Y	Hofstee Values Diff
15	6	45	39	28		
16	10	42	33	23		
17	13	40	27	19		
18	16	37	22	16		
19	21	35	14	10		
20	29	33	4	4		
21	40	30	-10	11	30	-11
22	50	28	-22	21	27	-23
23	61	25	-36	33	25	-37
24	72	23	-49	43		
25	78	21	-57	49		
26	86	18	-67	58		
27	89	16	-74	62		
28	94	13	-81	67		
29	96	11	-86	70		
30	98	9	-89	72		

References

Bejar, I.I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310.

Beuk, C.H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 21, 147-152.

DeGruijter, D.N.M. (1985). Compromise models for establishing examination standards. Journal of Educational Measurement, 22, 263-269.

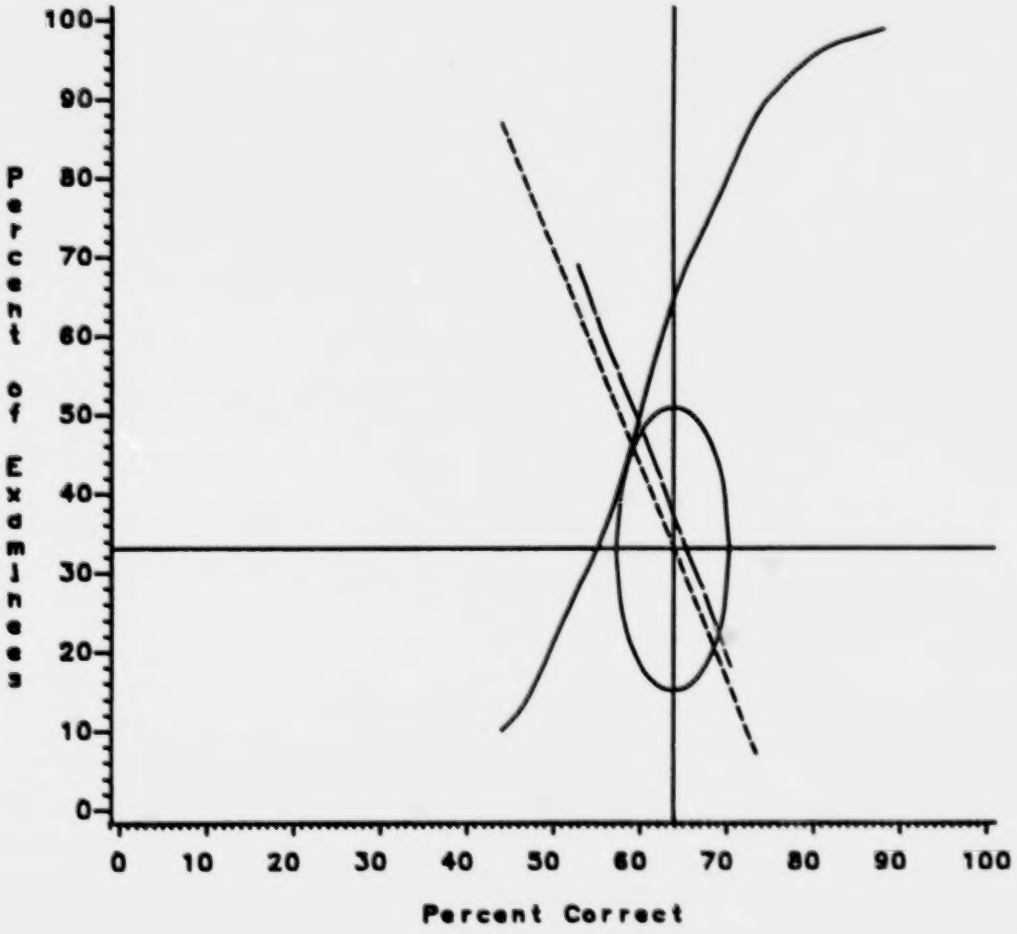
Lord, F.M. (1984) Standard errors of measurement at different ability levels. Research Report 84-8, Princeton, NJ: Educational Testing Service.

Lorge, I. & Kruglov, L.K. (1953) The improvement of estimates of test development. Educational and Psychological Measurement, 13, 34-46.

Livingston, S.A. & Zieky, M.J. (1982) Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests, Princeton, NJ: Educational Testing Service.

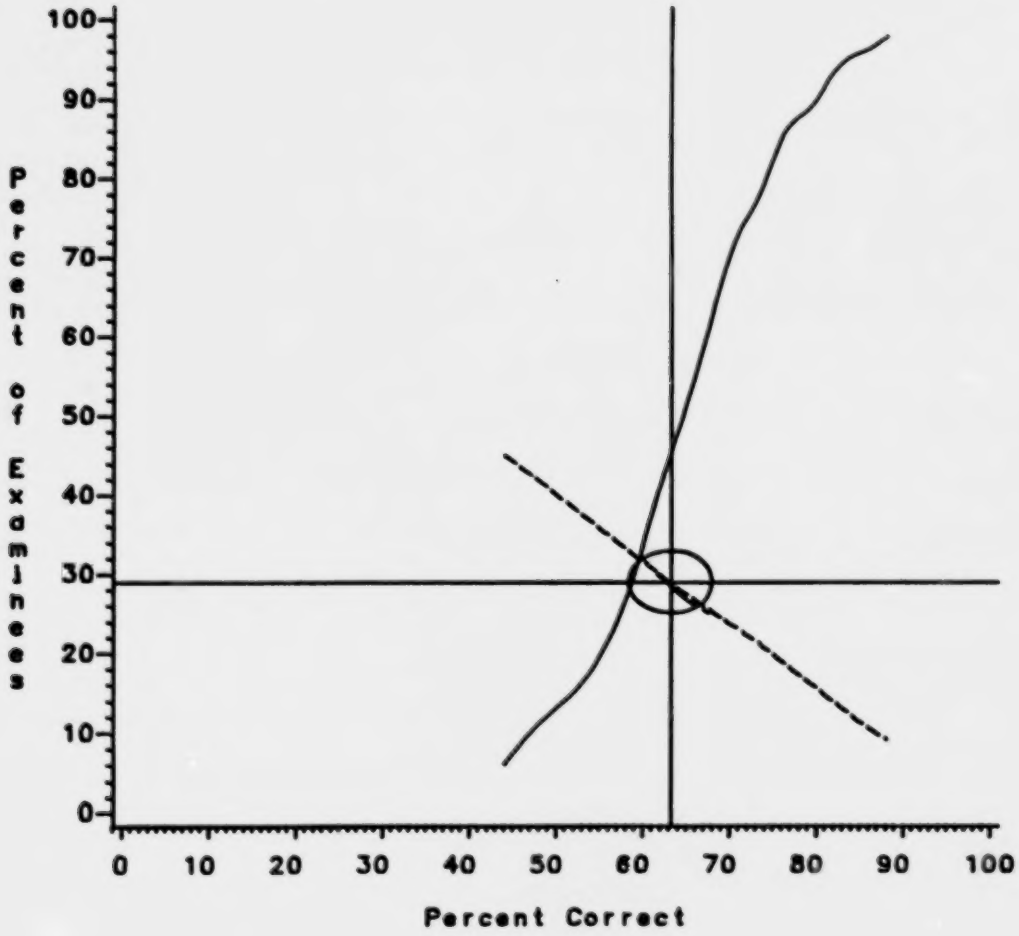
Thorndike, R.L. (1982) Item and score conversion by pooled judgment. In Holland, P.W. & Rubin, D.B. (eds.), Test Equating, New York: Academic Press, 309-317.

Figure 1: Compromise Results for Test 1
Beuk, Hofstee, & DeGruijter



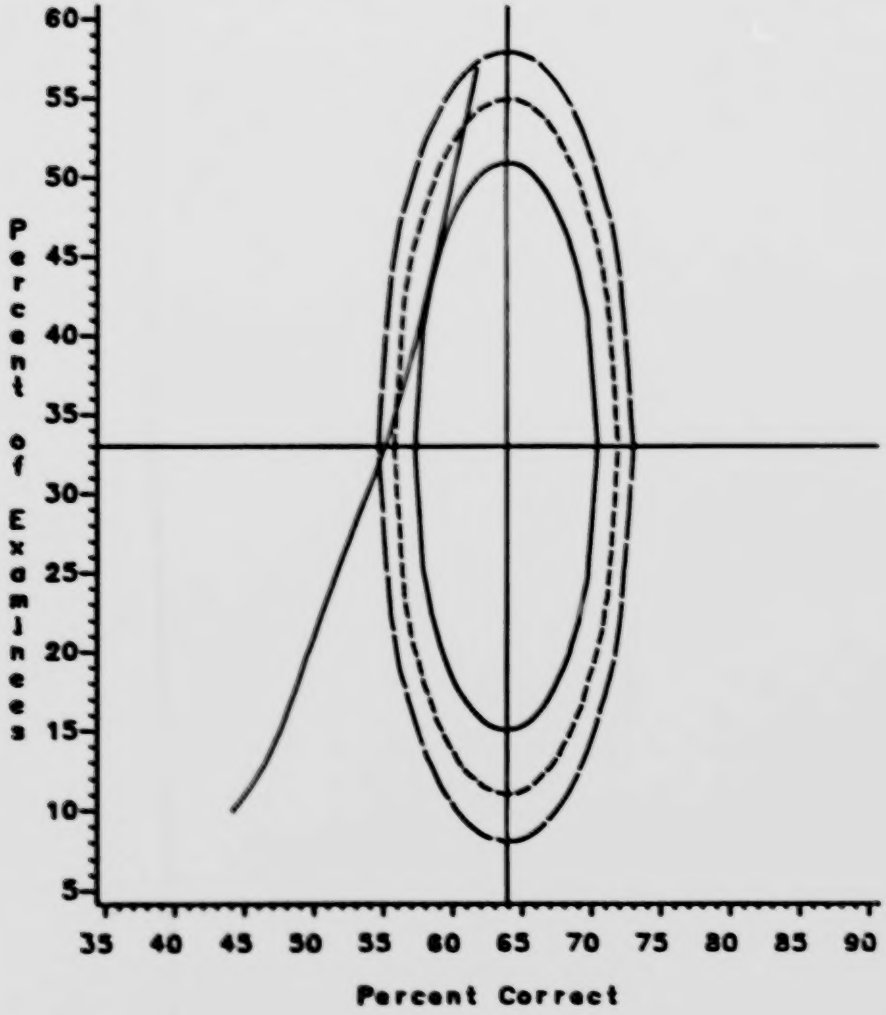
Beuk = Short Dashed Line
DeGruijter = Solid Ellipse
Hofstee = Long Dashed Line
Cumulative % Below = Solid Ogive

Figure 2: Compromise Results for Test 2
Beuk, Hofstee, & DeGruijter



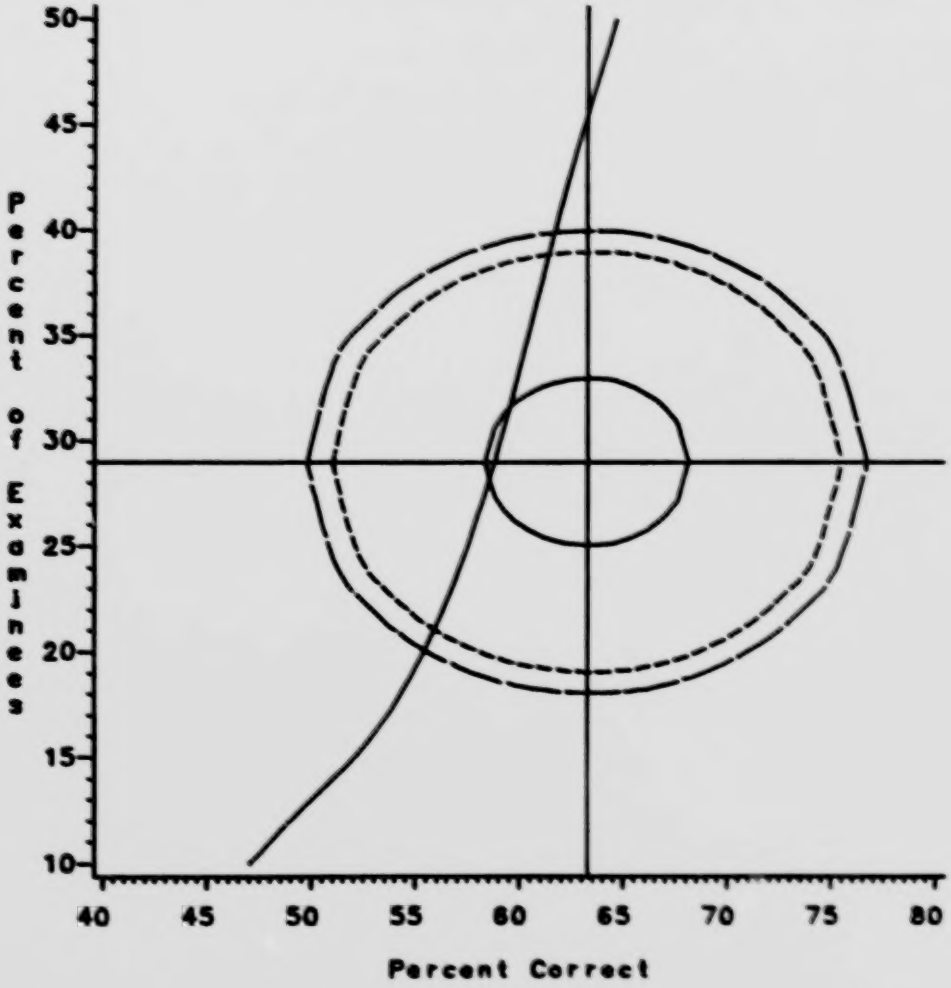
Beuk = Short Dashed Line
DeGruijter = Solid Ellipse
Hofstee = Long Dashed Line
Cumulative % Below = Solid Ogive

Figure 3: DeGruijter Ellipses for Test 1
Raw Cuts 19, 20, & 21



Score 19 = Short Dashed Ellipse
Score 20 = Solid Ellipse
Score 21 = Long Dashed Ellipse
Cumulative % Below = Solid Ogive

Figure 4: DeGruijter Ellipses for Test 2
Raw Cuts 19, 20, & 21



Score 19 = Short Dashed Ellipse
Score 20 = Solid Ellipse
Score 21 = Long Dashed Ellipse
Cumulative % Below = Solid Ogive